Using multimodal speech production data to evaluate articulatory animation for audiovisual speech synthesis

Ingmar Steiner*
University College Dublin
& Trinity College Dublin

Korin Richmond†
University of Edinburgh

Slim Ouni‡
Université de Lorraine
LORIA, UMR 7503

1 Introduction

The importance of modeling speech articulation for high-quality audiovisual (AV) speech synthesis is widely acknowledged. Nevertheless, while state-of-the-art, data-driven approaches to facial animation can make use of sophisticated motion capture techniques, the animation of the intraoral articulators (viz. the tongue, jaw, and velum) typically makes use of simple rules or viseme morphing, in stark contrast to the otherwise high quality of facial modeling. Using appropriate speech production data could significantly improve the quality of articulatory animation for AV synthesis.

2 Articulatory animation

To complement a purely data-driven AV synthesizer employing bimodal unit-selection [Musti et al. 2011], we have implemented a framework for articulatory animation [Steiner and Ouni 2012] using motion capture of the hidden articulators obtained through electromagnetic articulography (EMA) [Hoole and Zierdt 2010]. One component of this framework compiles an animated 3D model of the tongue and teeth as an asset usable by downstream components or an external 3D graphics engine. This is achieved by rigging static meshes with a pseudo-skeletal armature, which is in turn driven by the EMA data through inverse kinematics (IK). Subjectively, we find the resulting animation to be both plausible and convincing. However, this has not yet been formally evaluated, and so the motivation for the present paper is to conduct an objective analysis.

3 Multimodal speech production data

The `mngu0` articulatory corpus[1] contains a large set of 3D EMA data [Richmond et al. 2011] from a male speaker of British English, as well as volumetric magnetic resonance imaging (MRI) scans of that speaker's vocal tract during sustained speech production [Steiner et al. 2012]. Using the articulatory animation framework, static meshes of dental cast scans and the tongue (extracted from the MRI subset of the `mngu0` corpus) can be animated using motion capture data from the EMA subset, providing a means to evaluate the synthesized animation on the generated model (Figure 1).

4 Evaluation

In order to analyze the degree to which the animated articulators match the shape and movements captured by the natural speech production data, several approaches are described.

- The positions and orientations of the IK targets are dumped to data files in a format compatible with that of the 3D articulograph. This allows visualization and direct comparison of the animation with the original EMA data, using external analysis software.
- The distances of the EMA-controlled IK targets to the surfaces of the animated articulators should ideally remain close to zero during deformation. Likewise, there should be collision with a reconstructed palate surface, but no penetration.
- A tongue mesh extracted from a volumetric MRI scan in the `mngu0` data, when deformed to a pose corresponding to a given phoneme, should assume a shape closely resembling the vocal tract configuration in the corresponding volumetric scan.

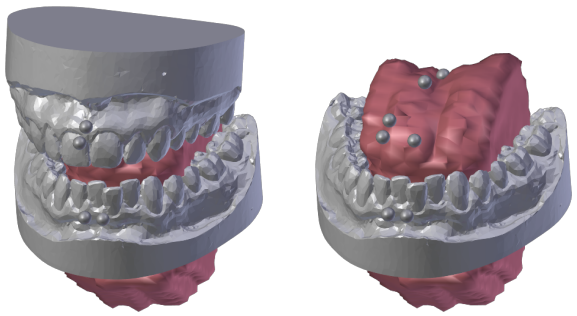

Figure 1: *Animated articulatory model in bind pose, with and without maxilla; EMA coils rendered as spheres.*

These evaluation approaches are implemented as unit and integration tests in the corresponding phases of the model compiler's build lifecycle, automatically producing appropriate reports by which the naturalness of the articulatory animation may be assessed.

References

HOOLE, P., AND ZIERDT, A. 2010. Five-dimensional articulography. In *Speech Motor Control: New developments in basic and applied research*, B. Maassen and P. van Lieshout, Eds. Oxford University Press, 331–349.

MUSTI, U., COLOTTE, V., TOUTIOS, A., AND OUNI, S. 2011. Introducing visual target cost within an acoustic-visual unit-selection speech synthesizer. In *Proc. 10th International Conference on Auditory-Visual Speech Processing (AVSP)*, 49–55.

RICHMOND, K., HOOLE, P., AND KING, S. 2011. Announcing the electromagnetic articulography (day 1) subset of the mngu0 articulatory corpus. In *Proc. Interspeech*, 1505–1508.

STEINER, I., AND OUNI, S. 2012. Artimate: an articulatory animation framework for audiovisual speech synthesis. In *Proc. ISCA Workshop on Innovation and Applications in Speech Technology*.

STEINER, I., RICHMOND, K., MARSHALL, I., AND GRAY, C. D. 2012. The magnetic resonance imaging subset of the mngu0 articulatory corpus. *Journal of the Acoustical Society of America 131*, 2 (Feb.), 106–111.

*`ingmar.steiner@ucd.ie`
†`korin@cstr.ed.ac.uk`
‡`slim.ouni@loria.fr`

[1] freely available for research purposes from `http://mngu0.org/`